sciendo

DOI: 10.2478/bile-2019-0007

Biometrical Letters
Vol. 56 (2019), No. 2, 117-138

Gene selection ensembles and classifier ensembles for medical diagnosis

Małgorzata Ćwiklińska-Jurkowska

Department of Theoretical Foundations of Biomedical Sciences and Medical Informatics, Collegium Medicum in Bydgoszcz, Nicolaus Copernicus University in Toruń, Jagiellońska 13-15, 85-067 Bydgoszcz, Poland,
e-mail: mjurkowska@cm.umk.pl

SUMMARY

The usefulness of combining methods is examined using the example of microarray cancer data sets, where expression levels of huge numbers of genes are reported. Problems of discrimination into two groups are examined on three data sets relating to the expression of huge numbers of genes. For the three examined microarray data sets, the cross-validation errors evaluated on the remaining half of the whole data set, not used earlier for the selection of genes, were used as measures of classifier performance. Common single procedures for the selection of genes—Prediction Analysis of Microarrays (PAM) and Significance Analysis of Microarrays (SAM)—were compared with the fusion of eight selection procedures, or of a smaller subset of five of them, excluding SAM or PAM. Merging five or eight selection methods gave similar results. Based on the misclassification rates for the three examined microarray data sets, for any examined ensemble of classifiers, the combining of gene selection methods was not superior to single PAM or SAM selection for two of the examined data sets. Additionally, the procedure of heterogeneous combining of five base classifiers—k-nearest neighbors, SVM linear and SVM radial with parameter c=1, shrunken centroids regularized classifier (SCRDA) and nearest mean classifier—proved to significantly outperform resampling classifiers such as bagging decision trees. Heterogeneously combined classifiers also outperformed double bagging for some ranges of gene numbers and data sets, but merging is generally not superior to random forests. The preliminary step of combining gene rankings was generally not essential for the performance for either heterogeneously or homogeneously combined classifiers.

Key words: combined methods, discriminant analysis, gene selection

1. Introduction

Seeking biomarkers to diagnose cancers at a very early stage, when it is easier to treat patients successfully, is a very important current research problem. Many efforts have been made in this area (see, for example, Cohen et al., 2017; Cohen et al., 2018). Assessment of selection is performed in the next classification stage, e.g. using artificial intelligence or machine or statistical learning.

Because of the high number of investigated genes in one microarray, the pre-selection of features for inclusion in the classification rule is essential. In high-dimensional problems we encounter the need to lower the number of variables to reduce the information noise. Often, only a few tens of genes are really active. The problem of selecting discriminating genes or proteins from microarrays is important in obtaining diagnostic markers. Some authors refer to such problems as "seeking a needle in a haystack" (for example Kumarasinghe et al., 2016). The remaining genes are not important for improvement of the discriminant procedure. In supervised classification problems, the variables with the greatest discriminant power are sought. Next, supervised classification is often used as an additional criterion to verify the correctness of a selected set of variables. This is done by the estimation of an error for a new independent sample by cross-validation or bootstrap procedures.

For classification issues involving microarray data sets, bagging (proposed by Breiman, 1998) or boosting combined classifiers are typically applied (e.g. Dettling et al., 2003, and see review by Boulesteix et al., 2008). Ensembles of classifiers based on resampling, like bagging, boosting or random subspace classifiers, may improve stability (as described, for example, in Skurichina et al., 2002). These methods may be called families of classifiers. Families are considered as homogeneous ensembles, because all base classifiers that are merged in one decision are of the same type, but the classifiers are created on slightly different subsets. Examples are random subset classifiers (the most

popular being random forests) and bagging classifiers based on the bootstrapping technique and classifier aggregation.

Combining procedures have recently often been applied in biomedical and bioinformatics applications (for example Cohen et al., 2018; Cohen et al., 2017; Dettling et al., 2003; Dettling, 2004; van Sanden et al., 2008). An ensemble of selection methods may lead to benefits in the final combined ranking of the most discriminating genes, because merging may incorporate into the joint ranking the different benefits of various methods. Here, the usefulness of combining was examined for a dimension reduction task and for a classifier construction stage, and for both of these jointly.

2. Methods

The method of manipulating the three examined data sets of gene expression levels is the same, although the sets contain different numbers of genes and cases. The three studied data sets are applied in important biomedical classification problems: differentiating between colon cancer and normal colon tissue samples (the Colon data set), discrimination between acute lymphoblastic leukemia (ALL) and acute myelogenous leukemia (AML) (the Leukemia data set) and identification of genotoxic and non-genotoxic compounds among carcinogenic chemical compounds (the GTX data set).

The GTX data set, described in van Delft et al. (2005), consists of 596 gene expression profiles from treated HepG2 cells, serving to discriminate genotoxic from non-genotoxic carcinogens. GTX contains microarrays for 24 genotoxic and 20 non-genotoxic chemical compounds.

The Leukemia data set (Golub et al., 1999) contains expression levels of 3571 genes. The number of samples is 72, all acute leukemia patients, with either acute lymphoblastic leukemia (ALL, 47) or acute myelogenous leukemia (AML, 25).

A detailed description of the handling of data sets with general denotations, but also with concrete numbers of genes and cases, is given only for the Colon data set, because for GTX and Leukemia the procedures are the same.

In the Colon data set (Alon et al., 1999), containing expression levels of v=2000 genes (variables, numbered from 1 to v) with c=62 cases (patients), 40 tissues out of 62 are colon tumor tissues and 22 are normal. The data set was standardized by subtracting a gene average over all samples and dividing by the standard deviation. The whole data set was randomly divided into two subsets S and D with similar numbers of patients, where S is used for selection of the most discriminating genes, and D for assessment of the discrimination. Thus S, a subset of the Colon data set, contains all v=2000 genes and n_1=32 patients, where 20 are tumor cases. For each discrimination problem, the first subset S (of n_1=32 patients and v genes) was used to select the most discriminating genes. The rankings of genes by importance are obtained by single methods (like PAM or SAM) or combined ranking methods. The subset of g genes consisted of the genes from first (the most discriminating) to g-th in the sequence of importance.

Subsequent subsets of selected genes S_g ($S_g \subseteq \{1,2,\ldots,100, v\}$) are increasing, and for single selection procedures include all genes selected in the previous set, i.e. $S_{g1} \subseteq S_{g2}$ for $g_1 \leq g_2$. Thus, after the selection, the variables are arranged according to the selection criterion. The discriminant power of the variable set S_g is associated with the single selection criterion. Different methods of dimension reduction were applied and are shown in Table 1 (in the first eight rows). The methods are described in detail in various publications. The single selection methods applied in the investigation were SAM (Significance Analysis of Microarrays, proposed by Tusher et al., 2001) and PAM (Prediction Analysis of Microarrays, proposed by Tibshirani et al., 2002). Additionally, ensembles of selections excluding SAM and PAM were constructed and compared with single SAM and PAM. The sequence of genes obtained by the merged procedures also indicates decreasing importance. The genes are arranged by ranks (between *1* and v), and according to the obtained sequence, subsets of genes are then used for evaluation of the examined classifiers' errors.

To find relevant genes by another method than the popular SAM and PAM, variable selections of different types were considered for merging into one ranking. The base rankings used for combination are obtained according to

Table 1 Single and combined gene selection methods applied

Identifier	Name	Description
1	Gini	Gini index
2	PermutAdjPminP	permutation adjusted Welch t-test, based on minimum adjusted p (Ge et al., 2003; Westfall et al., 2001)
3	PermutAdjPminPWilcox	permutation adjusted Wilcoxon test based on minimum adjusted p (Ge et al., 2003; Westfall et al., 2001)
4	PermutAdjPmaxT	permutation adjusted based on Welch t-test (Ge et al., 2003; Westfall et al., 2001)
5	PermutAdjPmaxWilcox	permutation adjusted based on Wilcoxon test (Ge et al., 2003; Westfall et al., 2001)
6	BH_AdjT	Wilcoxon test with Benjamini–Hochberg correction for multiple t-tests (Benjamini et al., 1995)
7	BH_AdjWilcox	Wilcoxon test with adjusted Benjamini–Hochberg correction for multiple Wilcoxon tests (Benjamini et al., 1995)
8	BetweenWithinRatio	ANOVA test, equivalent to t in 2 groups
9	SAM	Significance Analysis of Microarrays (Tusher et al., 2001)
10	PAM	Prediction Analysis of Microarrays (Nearest Shrunken Centroids, Tibshirani et al., 2002)
11	Comb.Sel1	Methods 1–5 in this table, combined
12	Comb.Sel2	Methods 1–8 in this table, combined

the Gini impurity measure, the between-groups-to-within-groups diversity ratio (BetweenWithinRatio), and the T and Wilcoxon tests with Benjamini–Hochberg adjustment for multiplicity (described e.g. in Benjamini et al., 1995; denoted here by BH_AdjT and BH_AdjWilcox respectively). Adjusted p-values for multiple testing procedures were also applied in permutation tests, and permutation Welch T and Wilcoxon rank-sum tests are incorporated into the combined arrangement (following Ge et al., 2003; Westfall et al., 1993; Westfall et al., 2001). The permutation algorithm for the maxT and minP procedures was introduced by Ge et al. (2003), and according to the description of the method, the base selection methods are denoted as PermutAdjPmaxT, PermutAdjPminP, PermutAdjPmaxWilcox and PermutAdjPminPWilcox. Here "Permut" denotes permutation, "minP" signifies the criterion of minimum p-value, "max" indicates the maximum statistic value (Welch T or Wilcoxon respectively), and "AdjP" signifies adjustment for multiplicity tests.

Two subsets of single selection procedures R_1 (1,…,5) and R_2 (1,…,8) are merged into combined rankings (denoted as Comb.Sel1 and Comb.Sel2 in Table 1). The subset R_1 is based on five selection methods: Gini, PermutAdjPmaxT, PermutAdjPminP, PermutAdjPmaxWilcox and PermutAdjPminPWilcox. The subset R_2 contains the following eight methods to combine gene rankings: BetweenWithinRatio, Gini, PermutAdjPmaxT, PermutAdjPminP, PermutAdjPmaxWilcox, PermutAdjPminPWilcox, BH_AdjT and BH_AdjWilcox. Thus, it includes the five earlier mentioned selection procedures ($R_1 \subseteq R_2$). For each gene number g, the sum of ranks obtained from the rankings in R_i (i=1,2) is calculated, i.e. the rank of 5 or 8 combined methods

$$\sum_{j=1}^{\overline{\overline{R_i}}} r_{g,j}$$

is the criterion for arrangement into the combined ranking. The ensemble variable selection procedures introduces a weight ranking of genes using the base selection methods' ranking in such a way that higher combined importance is assigned to a gene that occurs prior to others in most of the combined base rankings. Such merged rankings of five (Comb.Sel1) or eight (Comb.Sel2) procedures were investigated to compare them with the single SAM and PAM procedures.

To reduce the complexity of discrimination assessment and the calculation time, successive subsets of genes of increasing size were chosen as 15 increasing values from 1 to 100 (2, 5, 10, 15, 20, 25, 30, 35, 40, 45, 50, 55, 60, 70 and 100), and additionally v=2000 was applied for some classifiers for which this was possible. Thus, various discriminant methods are compared for an ascending number of genes: $g \in \{2, 5, 10, 15, 20, 25, 30, 35, 40, 45, 50, 55, 60, 70, 100, v\}$.

Because the number of considered chosen variables can even exceed 100, the classical discrimination fails for the applied CV procedure constructed from the set D. Thus, discriminant functions viable for high dimensionality were considered for merging. Various procedures have been discussed as alternatives to classical discriminant analysis. Some of these are shrunken centroids

regularized discriminant analysis (SCRDA; Guo et al., 2005), k-nearest neighbors discrimination (kNN), uncorrelated linear discrimination and Support Vector Machines (Cortes et al., 1995).

One of the base classifiers is Support Vector Machines (SVM). It provides an optimally separating hyperplane in the sense that the margin between two groups is maximized. In this work, SVM with linear kernel and with radial kernel were incorporated into the classifier ensemble; the default penalty parameter c=1 was applied. Also the k-nearest neighbor (kNN) discriminant function was included in the ensemble. In kNN, the number of neighbors is optimized according to the cross-validation error. Shrunken centroids regularized discriminant analysis (SCRDA) was also added to the ensemble. The last, fifth base classifier incorporated into the ensemble is a special case of linear diagonal (uncorrelated) discriminant function, assuming equal variances of genes (i.e. the nearest mean classifier). The ensemble classification method uses the classifier merging the results of the base classifiers, and we may expect this to be beneficial, because the constituent methods differ in methodology. Thus, the final joint classifier is heterogeneous, and is named here Heterogeneous Merge and denoted as HeterMerge2 (independently of the single or combined preliminary selection procedure). Majority voting was used for joint decisions.

Apart from the aforementioned single classifiers, resampling methods are also examined. Researchers in classification tend to combine procedures based on similar types or different base classifiers. Specifically, considerable attention has been paid recently to families of classifiers originating from two ideas: bootstrap aggregation and boosting. In the current work, the aggregation of classifiers constructed based on data set bootstrapping was applied. This class of bootstrap aggregation procedures includes typical bagging decision trees, modified double bagging with LDA (linear discriminant analysis; Hothorn et al., 2003) and double bagging with SLDA (stabilized LDA, proposed by Kropf, 2000). Double bagging combines LDA (or SLDA) and bagging classification trees.

Also, the random subspace method uses bootstrap aggregation with an additional step of random selection of variables in each loop. The application of decision trees gives the random forests procedure, introduced by Breiman (2001). In the current work a random forests classifier is applied, and congruently, trees are also investigated as base classifiers in other bootstrap aggregation procedures such as bagging, LDA double bagging and SLDA double bagging.

The generalization properties of selected subsets of genes were examined for increasing subsets of g genes, where g is in the set I={1, 2, 5, 10, 15, 20, 25, 30, 35, 40, 50, 60, 70, 100}. Each time, in the loop increasing the number of selected genes g (from set I), the submatrix M_g (with n_2 cases and g genes) of the expression level matrix M (with n_2 cases and v genes) is applied. It represents g previously chosen informative genes in the second independent set D. This matrix M_g can be treated as a g-dimensional representation of set D. Let us denote this set as D_g. For each $g \in$ I , the D_g was used for discrimination evaluation of n_2=30 patients (e.g. for the Colon data set, including 20 tumor patients and 10 normal cases).

Assessment of the performance of a classification method is performed by cross-validation with c=10 folds. Namely, for each number of genes g, the set D_g is randomly divided into c=10 disjoint subsets of approximately equal sizes: $D_{g,1}$, …, $D_{g,c}$. For each classified g-dimensional case x from the subset $D_{g,k}$, the classifier is built on $D_g \backslash D_{g,k}$ (for CV loop number $k=1,\ldots,c$).

Cross-validation errors $e_{g,k}$, evaluated on the subset $D_{g,k}$ of the data set D_g, were calculated for $k=1,\ldots,c$. Next the $e_{g,k}$ were averaged over c cross-validation loops, i.e.

$$e_g = \frac{1}{c}\sum_{k=1}^{c} e_{g,k}$$

For each g from set I, e_g was the criterion for comparisons of the correctness of selection procedures and classifiers. Additionally, identification of the optimal number of genes g_0 or ranges of gene numbers is possible based on the e_g errors.

To combine five constituent classifiers with different properties into one heterogeneous ensemble, the crisp results of the component classifiers $d_1,d_2,\ldots,d_5$ are subjected to voting (a crisp result indicates only the identifier of the recognized class: 1 or 2). The number of decisions $d_m=y$ classifying the object in the class y (y=1,2, m=1,…,5) is the ranking for selection of this class. The decision of the ensemble is arg max $_{y=1,2}$ $\sum(d_m=y)$. The result is explicit: ties in voting are not possible, because the number of combined classifiers is not divisible by the number of possible classes.

Similarly to single classifiers, for heterogeneously combined classifiers, $\forall\ g \in I$, the set D_g was divided into c=10 cross-validation sub-samples and the c cross-validation errors of the compared discriminant functions were calculated. Then the c-fold cross-validation error is estimated as the average error over c loops.

The software used for the analysis is R packages with original scripts for combining selection, discrimination ensembles, combination of cross-validation of the D_g sets with discriminant procedures, and for making evaluation curves.

3. Results

Misclassification rates were presented for increasing numbers of variables up to 100, because the use of more than 100 genes proved not to be constructive. For successive subsets of genes, ranked by the examined combined selection method, misclassification rates of different classification methods were assessed. For sets with increasing numbers of variables, the cross-validation classification errors are calculated, so classifier evaluation curves may be plotted with these values represented on the horizontal axis. A comparison of combined classifier evaluation curves can be made on the basis of Figs. 1, 3 and 5, where the cross-validation technique of division of the data set into 10 folds was used to assess generalization errors. Methods of applied combined selection are presented on the lower subplots, where 10-CV errors of classification methods for successive subsets of genes, jointly ranked by combined selection methods, are given.

HeterMerge2 is represented by a dark solid line, and random forest by a light dashed line.

Comparing the subplots in Figure 1 for the Colon dataset with mean 10-CV errors with the lines obtained by adding and subtracting standard errors, we can observe that HeterMerge2 (solid line) outperforms typical bagging and double bagging. This is especially distinct for about 60 genes. For random forest a difference for more than 60 genes is also observed, but the advantage of HeterMerge2 appears to be smaller in comparison with bagging, doubleBagging LDA and doubleBaggingSLDA (Fig. 1). Thus, a benefit from the application of a heterogeneously merged classifier is apparent, in comparison with the other four (homogeneous) ensembles. The effect holds for both single SAM selection (Fig. 1, top), the PAM procedure (Fig. 1, middle) and for a combined ranking from Gini, PermutAdjPmaxT, PermutAdjPminP, PermutAdjPmaxWilcox and PermutAdjPminPWilcox (Fig. 1, bottom). Comparing the top and middle subplots in Figure 1, we observe very similar learning curves for single SAM and PAM. Additionally, a similar performance of classifiers is obtained for the selection ensemble Comb.Sel1 constructed from five procedures: Gini, PermutAdjPmaxT, PermutAdjPminP, PermutAdjPmaxWilcox and PermutAdjPminPWilc (Fig. 1, bottom).

For different gene selection techniques an apparent difference may be observed between heterogeneously merged and homogeneous classifiers based on bootstrap aggregating, for example bagging, LDA and SLDA bagging and random forest, for more than 50 genes.

In the next figures the performance of single and combined selections from a larger set (i.e. eight variable selection procedures; Figs. 2 and 6) or five selection procedures merged (Fig. 4) is presented using lines with areas representing the standard errors of cross-validation misclassification rates. This helps to assess the significance of differences in classification rates for heterogeneously combined classifiers and bagging trees. For complementary examination of the difference between Heterogeneous Merge classifiers (HeterMerge2) and bagging-type classifiers, areas with standard errors were examined for the single PAM and

SAM selection procedures (Figs. 2, 4, 6, upper and middle) and for combined selection rankings (Figs. 2, 4, 6, bottom). Hence, for the Colon data set the plots indicate significant differences between the heterogeneous ensemble HeterMerge2 and bagging trees for 50–80 genes (Fig. 2). Comparing the subplots of Figure 1 and Figure 2, we can conclude that there is no significant difference between single SAM (upper subplot) or PAM (middle subplot) selection and the combined selection procedure for all types of combined classifiers, i.e. those taken from bagging families (bagging trees, double bagging LDA, double bagging SLDA and random forests) and the heterogeneous ensemble. The considered ensembles of five selection procedures are not superior to the popular SAM and PAM methods of gene selection (Fig. 1, bottom subplot), and moreover the eight wider-ranging selection techniques used for joint ranking of variables are not superior according to the evaluation by CV classification errors (Fig. 2, bottom subplot). Five and eight aggregated selection methods (Comb.Sel1 and Comb.Sel2 respectively) gave very similar learning curves, as can be seen by comparing the bottom subplots of Figs. 1 and 2. Comparing the learning curves from the bottom subplots of Figure 1 with those from the bottom of Figure 2, we can conclude that increasing the set of five joint selection rankings to eight selection procedures (i.e. adding three selection procedures to the ensemble: BetweenWithinRatio, BH_AdjT and BH_AdjWilcox) is not advantageous for combined gene rankings, either for the performance of a bagging-type classifier or for the performance of the heterogeneously combined classifier HeterMerge2.

Following on from the comparison of the upper and middle subplots of Figure 1, it is seen that there is no essential difference with SAM or PAM selection in the upper and middle subplots of Figure 2. The small differences between the results for families of classifiers (bagging type and random forests) visible in Figure 1 or Figure 2 originate from the non-deterministic bootstrap aggregation technique used in random forests and bagging-type classifiers. Further, the small difference between the same merged classifiers in the bottom subplots of Figs. 1 and 2 originates from the different procedures used

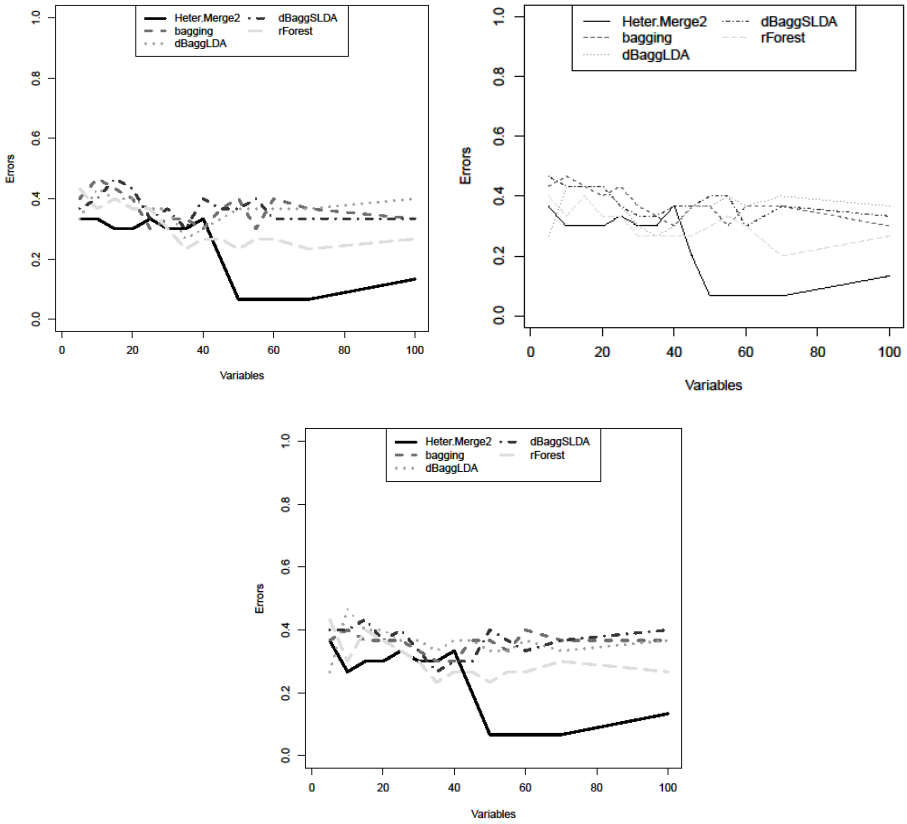

Figure 1. Ten-fold CV errors of homogeneous and heterogeneous combined classifiers after selection by SAM (upper plot), PAM (middle plot) and combined ranking from five procedures (Gini, PermutAdjPmaxT, PermutAdjPminP, PermutAdjPmaxWilcox, PermutAdjPminPWilcox; bottom plot) for the Colon data set

in preliminary joint rankings (the set of five merged selection procedures, Comb.Sel1; or that set enlarged to eight procedures, Comb.Sel2).

To analyze the GTX data set, Figs. 3 and 4 provide the possibility of making similar comparisons as for Figs. 1 and 2 respectively. For the Leukemia data set, corresponding evaluation curves are shown in Figs. 5 and 6. The meaning of evaluation curves, denotations of methods, colors and lines are the same as in the case of Figs. 1 and 2 for the Colon data set.

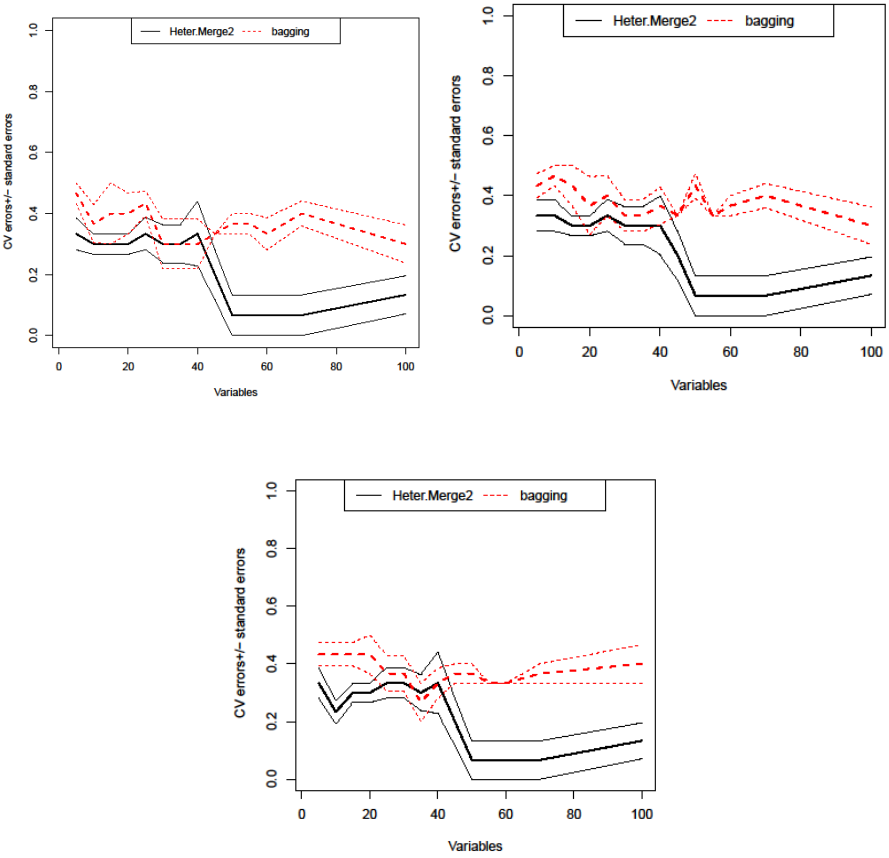

Figure 2. Ten-fold CV classification errors with standard errors of classification methods for the Colon data set: merged classifier (HeterMerge2) and bagging tree (100 loops), for successive subsets of genes ranked by single SAM selection (upper plot), PAM (middle plot), and by a combined ranking from eight selection procedures, Comb.Sel2 (bottom plot)

The three subplots in Figure 3 provide the possibility of comparing combined classifiers based on different selection methods for the GTX data set. Between 22 and 52 genes, HeterMerge2 (based on voting by k-nearest neighbor, regularized SCRDA classifier, nearest mean classifier, linear SVM and radial SVM) outperforms all of the other examined combined methods if the merged preliminary ranking (Comb.Sel1) from five procedures is applied (bottom subplot

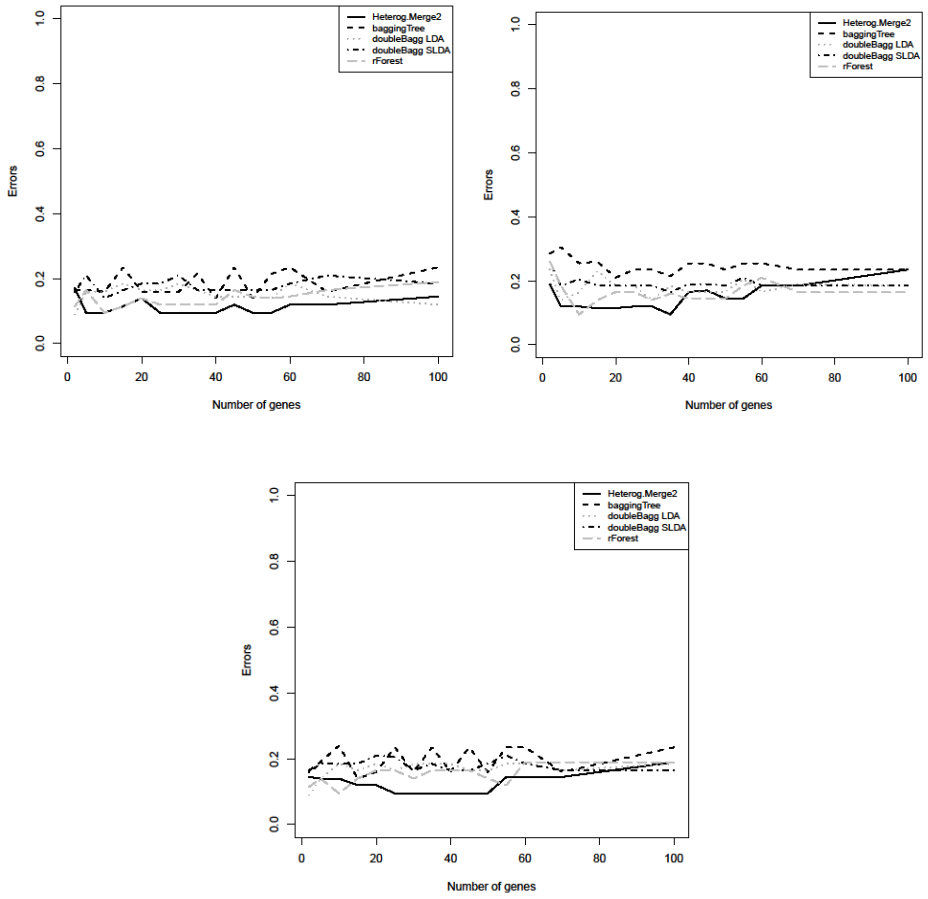

Figure 3. Ten-fold CV errors of homogeneous and heterogeneous combined classifiers after selection by SAM (upper plot), PAM (middle plot) and combined ranking from five procedures (Gini, PermutAdjPmaxT, PermutAdjPminP, PermutAdjPmaxWilcox, PermutAdjPminPWilcox; bottom plot) for the GTX data set

of Fig. 3). However, application of HeterMerge2 after single SAM and PAM selections only slightly outperformed the other combined classifiers (upper and middle subplots of Fig. 3).

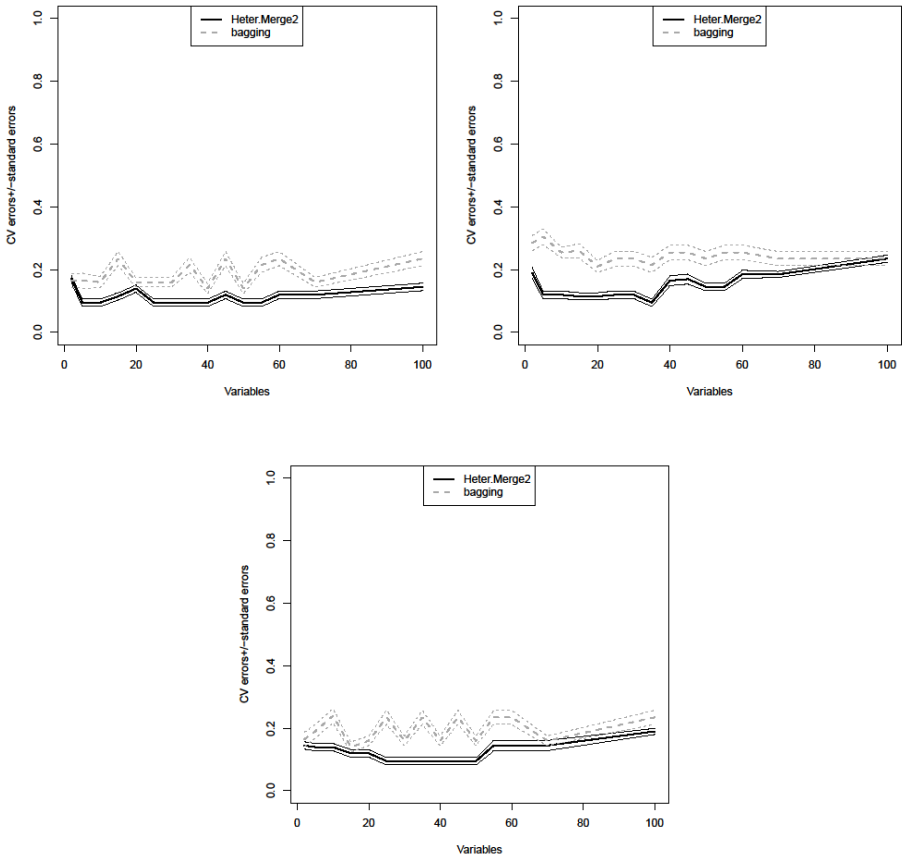

Figure 4. Ten-fold CV classification errors with standard errors of classification methods for the GTX data set: merged classifier (HeterMerge2) and bagging tree (100 loops), for successive subsets of genes ranked by single SAM selection (upper plot), PAM (middle plot) and by combined ranking from five procedures: Gini, PermutAdjPmaxT, PermutAdjPminP, PermutAdjPmaxWilcox, PermutAdjPminPWilcox (bottom plot)

Figure 4 presents three plots of estimated generalized classification errors for the GTX data set for HeterMerge2, i.e. a merged classifier with added and subtracted standard errors. These lines are visible as dark solid lines. Dashed lines represent the results of bagging tree (100 loops). SAM, PAM and Comb.Sel1 are represented by the upper, middle and bottom subplots respectively. Significant

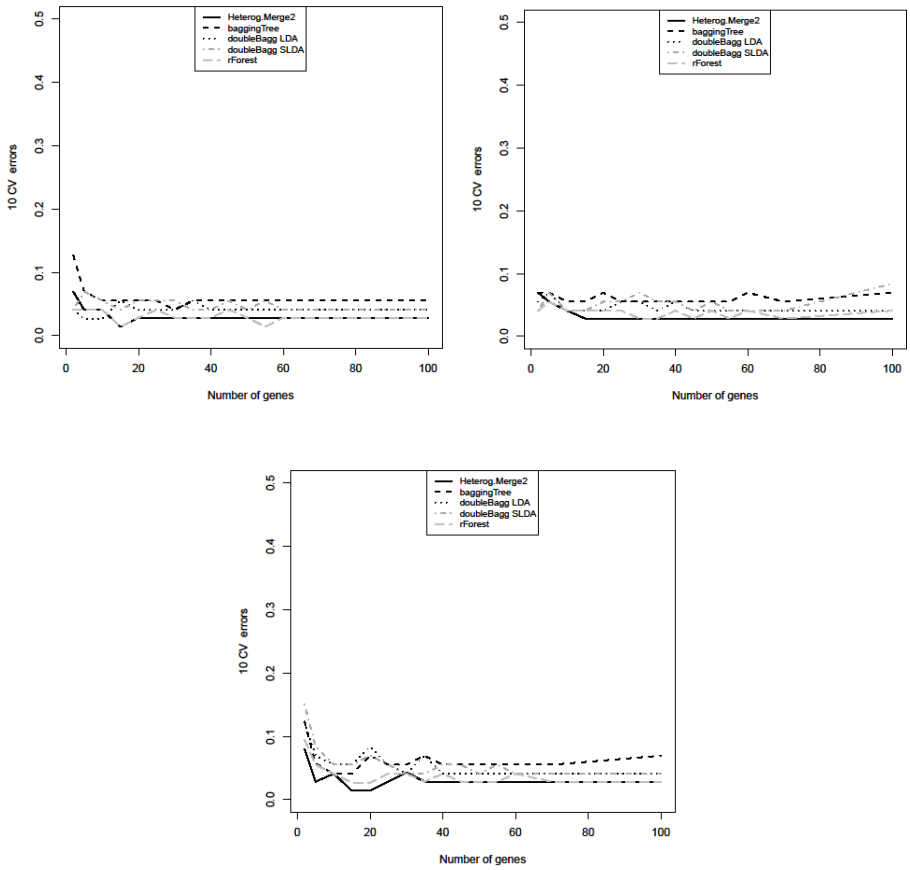

Figure 5. Ten-fold CV errors of homogeneous and heterogeneous combined classifiers after selection by SAM (upper plot), PAM (middle plot) and combined ranking from five procedures: Gini, PermutAdjPmaxT, PermutAdjPminP, PermutAdjPmaxWilcox, PermutAdj (bottom plot) for the Leukemia data set

differences between HeterMerge2 and bagging tree may be identified based on these plots, especially for smaller numbers of genes, e.g. for PAM selection and a number of genes not exceeding 35, or for Comb.Sel1 and a number of genes between 22 and 52.

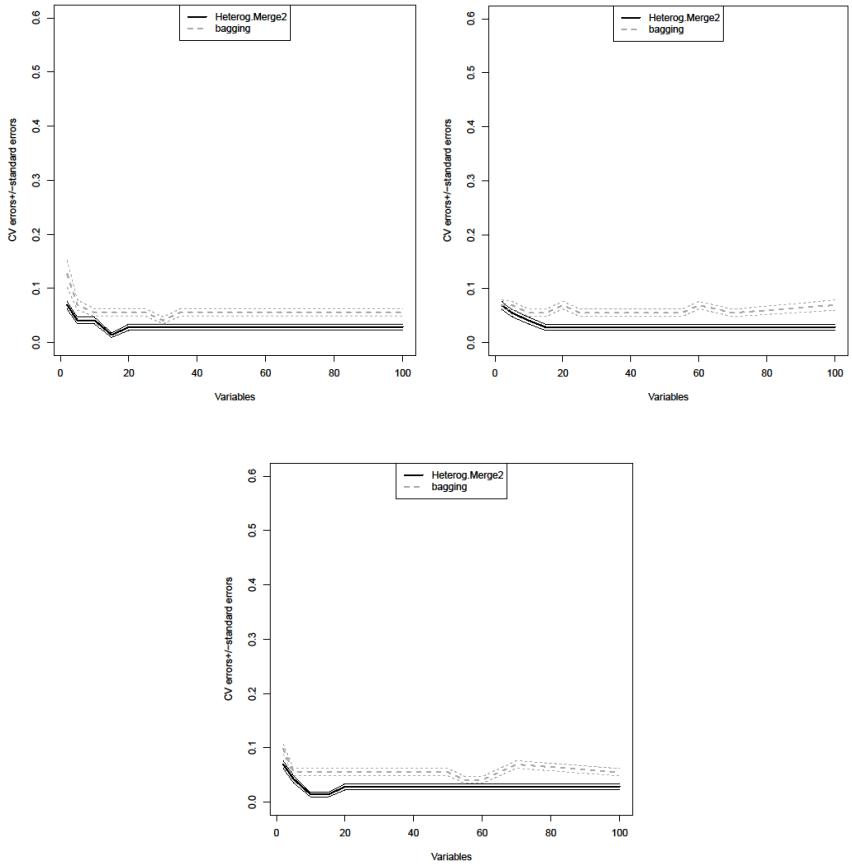

Figure 6. Ten-fold CV classification errors with standard errors of classification methods for the Leukemia data set: merged classifier (HeterMerge2) and bagging tree (100 loops), for successive subsets of genes ranked by single SAM selection (upper plot), PAM (middle plot) and by a combined ranking from eight selection procedures, Comb.Sel2 (bottom plot)

In a similar way to Figs. 1 and 3, the evaluation curves for the Leukemia data set are given in Figure 5. Also, analogously to Figs. 2 and 4, the comparison of classification errors of HeterMerge2 and bagging tree (100 loops) for the Leukemia data set is represented in Figure 6. For this specific data set the advantage of the HeterMerge2 classifier over all other combined classifiers is not apparent, because this set is less complicated for classification purposes, and the

classification errors are small for all selection methods and for all classifiers. However, from Figure 6 it may be concluded that, similarly to the other examined genomic data sets, the differences between the HeterMerge2 classifier and bagging trees with 100 loops are significant for some ranges or numbers of genes, where HeterMerge2 outperforms bagging. For SAM and Comb.Sel2 this is most apparent for numbers of genes below 20, while for PAM it occurs for greater than 20 variables.

On all plots the smallest benefit of the HeterMerge2 classifier is obtained with respect to random forests, visible as a light dashed line. Only for the GTX data set did HeterMerge2 with genes preliminarily selected by Comb.Sel1 outperform random forest for a range of genes between 22 and 52.

4. Discussion

The preceding step of combining rankings of variables, with either five or eight procedures, was not essential for the performance of all examined ensembles of classifiers. The different ensemble techniques examined for selection of variables are not essentially different according to evaluation by CV classification errors, and there is also no significant difference between them and SAM or PAM selection. For both types of combined classifiers (heterogeneous and homogenous) the performance is not increased after adding three selection rankings in the preliminary discrimination step, so enlarging the set of merged ranking procedures from five to eight is not beneficial. However, in the set of joint selection methods, some of them are based on the same ideas, such as permutation tests or the Benjamini–Hochberg correction for multiplicity. The question arises whether combining selections from a wider set of diverse procedures might improve the results for combined classifiers constructed on preliminarily selected variables.

The heterogeneous classifier HeterMerge2, obtained by majority voting based on the decisions of five constitutive discriminant procedures—k-nearest neighbor, regularized SCRDA classifier, Euclidean classifier (nearest mean),

linear SVM, radial SVM—was found to outperform some of the examined combined tree methods based on resampling of the examined data set, especially the homogeneous bagging trees classifier. For the Colon and GTX data sets, HeterMerge2 preceded by Comb.Sel1 outperforms all other examined combined methods if a merged preliminary ranking from five or eight procedures (Comb.Sel1 or Comb.Sel2) is applied. The base classifiers combined into the heterogeneous classifier are based on several different ideas from a wide range of pattern-recognition methodologies. The ensemble may take advantage of important and complementary or competitive benefits of the base methods.

The number of genes may be chosen as a trade-off between the size of the gene set and the goal of reducing error. The optimal gene subsets are indicated by the heterogeneous selection and classifier HeterMerge2 to be about 50–70 genes for the Colon data set, where the classification error maintains a stable level. For the GTX data set the optimal number of genes ranged between 22 and 52, while for the Leukemia data set the optimal range is between 10 and 20. Thus, the optimal numbers of genes are different in all of the examined classification methods for microarray data problems. Additionally, it should be noted that the performance of feature selection techniques in microarray classification scenarios is problem-dependent (as shown, for example, in Chai et al., 2004). This is concordant with our results, especially with regard to the different performance of classifiers obtained for the three examined datasets, even though the same selection methods and classifiers were examined.

5. Concluding remarks

Based on cross-validation errors, for the examined microarray data set, heterogeneous merging of classifiers performs significantly better than homogeneous ensembles like bagging, and for some ranges of genes also LDA double bagging and SLDA double bagging. Smaller 10-fold cross-validation errors, at least for some ranges of genes numbers, were achieved for a heterogeneous ensemble of regularized discriminant analysis (SCRDA), kNN,

linear and radial SVM and nearest mean classifiers than for homogeneous ensembles, independently of the prior variable selection method. The advantage may originate from the different ideas behind the merged constituent classifiers, because all of them have different properties and benefits.

However, the merging of gene rankings obtained from permutation procedures adjusted for multiplicity and Gini index, between-within group diversity, parametric and nonparametric testing with Benjamini–Hochberg adjustment for multiplicity was not beneficial over all examined families of tree classifiers, in terms of subsequent misclassification rates, for two of the three data sets: Colon and Leukemia.

The combining of procedures provided an advantage for the classifier building stage, but not for dimension reduction. The preliminary step of combining gene rankings was not essential for the performance for either heterogeneously or homogeneously combined classifiers.

The heterogeneously merged classifier combined from five discriminant methods, as well as with preliminarily combined five or eight selection procedures, proved superior to some families of classifiers based on bootstrap aggregating. The clearest advantage was obtained with respect to bagging trees, with a smaller or insignificant advantage over random forest. For two of the three examined data sets, preliminary selection by merging eight or five single arrangements of important genes was not beneficial, while for the GTX data set the preliminary selection ensemble proved advantageous prior to use of the heterogeneously merged classifier.

REFERENCES

Alon U., Barkai N., Notterman D.A., Gish K., Ybarra S., Mack D., Levine A.J. (1999): Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. Proc Natl Acad Sci USA. 96(12): 6745–50.

Benjamini Y, Hochberg Y. (1995): Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal Statistical Society Series B 57: 289–300.

Breiman L. (1996): Bagging predictions. Machine Learning 24 (2): 123–140.

Breiman L. (2001): Random Forests. Machine Learning 45: 5–32.
Boulesteix A.L., Strobl C., Augustin T., Daumer M. (2008): Evaluating Microarray-based Classifiers: An Overview. Cancer Inform. 6: 77–97.
Chai H., Domeniconi C. (2004): An Evaluation of Gene Selection Methods for Multi-class Microarray Data Classification. In: Proc. 2nd European Workshop on Data Mining and Text Mining in Bioinformatics, 3–10.
Cohen J.D., Li Y., Wang C., Thoburn B., Afsari L. et al. (2018): Detection and localization of surgically resectable cancers with a multi-analyte blood test. Science 10.1126/science.aar3247
Cohen J.D., Javed A.A, Li C., Thoburn, Wonga F., Tie J., Gibbs P. et al. (2017): Combined circulating tumor DNA and protein biomarker-based liquid biopsy for the earlier detection of pancreatic cancers. Proc Natl Acad Sci USA 114 (38): 10202–10207.
Cortes C., Vapnik V. (1995): Support-Vector Networks. Machine Learning 20: 273–297.
Dettling M., Bühlmann P. (2003): Boosting for tumor classification with gene expression data. Bioinformatics 19 (9): 1061–1069.
Dettling M. (2004): BagBoosting for tumor classification with gene expression data. Bioinformatics: 20: 3583–3593.
van Delft J.H., van Agen E., van Breda S.G., Herwijnen M.H., Staal Y.C., Kleinjans J.C. (2005): Comparison of supervised clustering methods to discriminate genotoxic from non-genotoxic carcinogens by gene expression profiling. Mutat Res, 575(1–2): 17–33.
Ge Y., Dudoit S., Speed T.P. (2003): Resampling-based multiple testing for microarray data analysis. January 2003. Technical Report 633.
Golub T.R., Slonim D.K., Tamayo P., Huard C., Gaasenbeek M., Mesirov J.P., Coller H., Loh M.L., Downing J.R., Caligiuri M.A., Bloomfield C.D., Lander E.S (1999): Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science, 286(5439): 531–537.
Guo Y., Hastie T., Tibshirani R. (2005): Regularized Discriminant Analysis and Its Application in Microarrays. Biostatistics, 1(1): 1–18.
Hothorn T., Lausen B. (2003): Double-bagging: combining classifiers by bootstrap aggregation. Pattern Recognition 36 (2): 1303–1309.
Kumarasinghe N., Tooney P.A., Schall U. (2012): Finding the needle in the haystack: A review of microarray gene expression research into schizophrenia. Australian & New Zealand Journal of Psychiatry 46 (7): 598–610.
van Sanden S., Lin D., Burzykowski T. (2008): Performance of gene selection and classification methods in a microarray setting: A simulation study. Communications in Statistics – Simulation and Computation 37(2): 409–424.
Skurichina M., Duin R.P.W. (2002): Bagging, Boosting and the Random Subspace Method for Linear Classifiers. Pattern Analysis & Applications 5:121–135.
Tibshirani R., Hastie T., Narasimhan B., Chu G. (2002): Diagnosis of multiple cancer types by shrunken centroids of gene expression. PNAS: 99: 6567–6572.
Tusher V., Tibshirani R., Chu G. (2001): Significance analysis of microarrays applied to the ionizing radiation response. PNAS 98: 5116–5121.

Westfall P.H., Zaykin D.V., Young S.S. (2001): Multiple tests for genetic effects in association studies. In: S. Looney (ed.), Methods in Molecular Biology 184: Biostatistical Methods, Humana Press, Toloway, NJ: 143–168.

Westfall P.H., Young S.S. (1993): Resampling-based multiple testing: Examples and methods for p-value adjustment. John Wiley & Sons.